A hybrid model for automatic emotion recognition in suicide notes

Biomedical Informatics Insights

ORIGINAL RESEARCH

A Hybrid Model for Automatic Emotion Recognition in Suicide Notes

Hui Yang[1], Alistair Willis[1], Anne de Roeck[1] and Bashar Nuseibeh[1,2]

[1]Department of Computing, Open University, Milton Keynes, United Kingdom. [2]Lero, University of Limerick, Limerick, Ireland. Corresponding author email: h.yang@open.ac.uk

Abstract: We describe the Open University team's submission to the 2011 i2b2/VA/Cincinnati Medical Natural Language Processing Challenge, Track 2 Shared Task for sentiment analysis in suicide notes. This Shared Task focused on the development of automatic systems that identify, at the sentence level, affective text of 15 specific emotions from suicide notes. We propose a hybrid model that incorporates a number of natural language processing techniques, including lexicon-based keyword spotting, CRF-based emotion cue identification, and machine learning-based emotion classification. The results generated by different techniques are integrated using different vote-based merging strategies. The automated system performed well against the manually-annotated gold standard, and achieved encouraging results with a micro-averaged F-measure score of 61.39% in textual emotion recognition, which was ranked 1st place out of 24 participant teams in this challenge. The results demonstrate that effective emotion recognition by an automated system is possible when a large annotated corpus is available.

Keywords: emotion recognition, keyword-based model, machine-learning-based model, hybrid model, result integration

Biomedical Informatics Insights 2012:5 (Suppl. 1) 17–30

doi: 10.4137/BII.S8948

This article is available from http://www.la-press.com.

Introduction

Recently, sentiment analysis has become an important line of research in computational linguistics. Emotion recognition is one type of sentiment analysis that focuses on identifying the emotion fragments (words, phrases, sentences) in free text. Automatic recognition of emotion from text presents an open research challenge due to the inherent ambiguity in emotion words and the rich use of emotion terminology in natural language. Various techniques have been proposed for textual emotion recognition. They include corpus-based techniques, such as using an emotion lexicon with weighted scores from training documents to build an emotion prediction model,[1] and machine learning-based approaches where an annotated corpus is used to train an emotion classifier,[2] as well as knowledge-based techniques that exploit linguistic rules based on the knowledge of sentence structures combined with several sentiment resources (eg, WordNet,[3] WordNet-Affect,[4] and SentiWordNet[5]) for emotion classification.[6]

This paper describes a system that uses a hybrid model to target for the emotion recognition task in the 2011 i2b2/VA/Cincinnati Medical Natural Language Processing Challenge. The system consists of a set of language models. These include a keyword spotting model with a pre-compiled list of weighted emotion terms trained from the training dataset, a Conditional Random Field (CRF)-based model for identifying emotion clues at the token level, and three different machine learning-based models, Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM), for emotion classification at the sentence level. The five language models compete with and complement one another in order to detect affective text of 15 emotions in a set of suicide notes.

Emotion Dataset for the Sentiment Analysis Task

The objective of the sentiment analysis task for the 2011 i2b2/VA/Cincinnati Challenge is to annotate, at the sentence level, the text in suicide notes with 15 specified emotion classes. We have grouped the 15 pre-specified classes into three sentiment polarity categories, *Positive Emotions*, *Negative Emotions*, and *Neutral Contexts* as follows:

- Negative Emotions (7): abuse, anger, blame, fear, guilt, hopelessness, sorrow
- Positive Emotions (6): forgiveness, happiness_peacefulness, hopefulness, love, pride, thankfulness
- Neutral Contexts (2): information, instructions

The dataset used for the sentiment analysis task consists of 900 suicide notes in which 600 annotated documents were released as the training data, and the rest of 300 unseen notes were used for the testing. The dataset was annotated by a team of over 160 volunteers who had lost a loved one to suicide. Each note was annotated by three different annotators. The inter-annotator agreement is approximately 0.535 at the token level and 0.546 at the sentence level. The statistical information about the dataset and associated sentiment polarities and individual emotions is shown in Tables 1, 2 and 3, respectively.

There are a number of interesting findings revealed in the training data:

a. Documents with annotated sentences: It is surprising that a few suicide notes in the dataset do not contain any emotion sentences. This situation can occur in two cases: one is because there is no positive, negative, or neutral emotion information presented in the note: another is due to the disagreement among the three annotators. It is also interesting that almost half of the sentences contain one or more of the emotion expressions of interest.
b. Sentiment polarities: Unlike other sentiment analysis work which tends to focus on the identification of positive and negative emotions, this challenge introduces *neutral contextual polarity*. This refers to text that describes instructions that the writer gave on what to do next or information about where things stand. The instances marked with the *information* or *instructions* label account

Table 1. The statistics information of the training and test data.

	Training	Test
#Document	600	300
#Document with annotated sentences	595	299
#Sentence	4633	2086
#Annotated sentences	2173 (46.9%)	1098 (52.6%)

Table 2. The distribution of emotion instances in different sentiment polarities.

	Training	Test
Sentiment		
Positive	483 (19.2%)	317 (24.9%)
Negative	924 (36.6%)	469 (36.9%)
Neutral	1115 (44.2%)	486 (38.2%)
Total	2522	1272

for about 44.2% in all of the annotated training instances, and 38.2% in the test instances, which suggests that neutral examples will play an important role in emotion analysis.

c. Emotion instance distribution: the number of emotion instances annotated with different emotion class labels varies widely, with ranges from just a few training instances for certain emotions such as *forgiveness* and *abuse*, to hundreds of examples for emotions like *Instructions* and *Hopelessness*. The emotions with scarce training instances cause great difficulty in emotion classification because there are not enough examples to train a ML-based emotion classifier to extract frequent emotion context patterns.

d. Sentences labeled with multiple emotion classes: Statistics on the annotated training dataset show that 13.9% of the annotated sentences contain affective text concerned with more than one emotion. Multi-emotion sentences are necessary when expressing complex feelings. Table 4 shows that some cases belong to typical cases of polarity shifting[7] where the sentiment of one sentence is changed from one polarity emotion (eg, *love*) to another polarity emotion (eg, *hopelessness*), whereas some cases are just the ones where two different emotions in the same sentiment polarity co-occur in the sentence, such as the co-occurred emotion pair, *hopelessness* and *guilt*. Table 5 illustrates some frequent emotion pairs co-occurring in the multi-emotion sentences. Such emotion co-occurrence information provides useful clues in emotion analysis. However, our approaches to emotion identification mainly focus on the recognition of individual emotions. Thus the analysis for emotion co-occurrence in the complicated sentences will not be utilized in current work but will be explored in the future work on multi-emotion sentence analysis.

Table 3. The distribution of emotion instances in individual emotions.

	Training	Test
Positive		
Forgiveness	6	8
Pride	15	9
Happiness_peacefulness	25	16
Hopefulness	47	38
Thankfulness	94	45
Love	296	201
Negative		
Abuse	9	5
Fear	25	13
Sorrow	51	34
Anger	69	26
Blame	107	45
Guilt	208	117
Hopelessness	455	229
Neutral		
Information	295	104
Instructions	820	382

Research Issues Related to Emotion Recognition

The analysis on the training data brings up several research issues that need to be addressed during the system development.

First, as Ortony[8] discussed, while some words (eg, *miserable*, *painful*) bear fairly unambiguous affective meaning, there are words that act only as *indirect* reference to emotion states, depending on the contexts in which they appear. Interestingly, we also found that, even words with the same sense can often evoke different emotions in certain contexts. Consider, for example, the underlined affect word *forgive* in the sentences (E1) and (E2). It evokes two different polarity emotions, *guilt* and *forgiveness* when it is followed by different pronouns. Therefore, detecting affective

Table 4. The distribution of sentiment polarity changes in the multi-emotion sentences.

Polarity_1	Polarity_2	#Sentence (training)	#Sentence (test)
Positive	Positive	14	16
Positive	Negative	89	66
Positive	Neutral	63	32
Negative	Negative	82	38
Negative	Neutral	94	34
Neutral	Neutral	67	15

Table 5. Part of the frequent co-occurred emotion pairs in the multi-emotion sentences.

Sentiment polarity	Emotion_1	Emotion_2	#Sentence (training)	#Sentence (test)
Positive/Positive	Love	Thankfulness	7	6
	Love	Hopefulness	3	2
Positive/Negative	Love	Hopelessness	31	16
	Love	Guilt	18	20
Positive/Neutral	Love	Instructions	40	23
	Thankfulness	Instructions	7	4
Negative/Negative	Hopelessness	Guilt	34	20
	Anger	Blame	10	4
Negative/Neutral	Hopelessness	Instructions	38	9
	Guilt	Instructions	16	15
Neutral/Neutral	Information	Instructions	67	15

text needs to consider the neighboring context of the affect word.

E1: *E1.Tell him to forgive me if I ever treated him bad.* [Emotion: *guilt]*

E2: *Tell him I forgive him for all my heart aches.* [Emotion: *forgiveness*]

Second, although the sentiment of many sentences is indicated by the presence of affect words, quite a number of sentences do not contain such words but convey affect through the underlying meaning. An example (E3), which does not contain an expected affect word, is given below. Automatically detecting such pragmatic information is a hard challenge, and the language models that rely on surface features of the sentences are very weak in detecting this kind of sentences with *implicit* emotion expressions.

E3: *I do n't know where she put my clothes from my dresser.* [Emotion: *anger*]

Third, as mentioned earlier, quite a number of sentences contain two or more emotion expressions. For example, in the sentence (E4), the first clause conveys a *fear* emotion through the verb phrase "*afraid of*", but the second clause conveys a *love* emotion by the verb "*love*". Because of the small number of multi-emotion instances, it is impractical to build multi-emotion classifiers to distinguish the multi-emotion sentences from the text. One feasible solution might be to build multiple binary classifiers, each of which is just targeted to one particular emotion. However, for a sentence-level binary emotion classifier, the text fragment depicting other emotions will become the noisy data, which is likely to degrade the accuracy of the classifier. Therefore, further fine-grained emotion analysis at the smaller text unit level (ie, emotion cues) is required. For example, emotion cues (eg, "*I am afraid of you*", "*I love you"*) that convey affective meaning with respect to a particular emotion needed to be separately annotated from the sentences. The annotation of emotion cues is discussed in a later section.

E4: *It is just that I am afraid of you both at times, but I love you both very much.* [Emotions: *fear, love*]

Fourth, we found that affective text of some emotions (eg, *hopelessness*) is sensitive to negation expressions. Certain phrases that contain negation words, eg, "*cant go on*", "*can't stand*", and "*can not take it any more*", intensify the emotion strength. Moreover, negation words sometimes can trigger the polarity shifting of an emotion, such as "*I do not blame him*". Therefore, it is necessary to incorporate negation detection into the identification of emotion expressions.

Fifth, while machine learning-based models may be capable of effectively classifying the emotions (eg, *love*, *hopelessness*, *guilt*, etc.) with a sufficient number of training instances, they do not work well on the emotions that have few training examples (eg, *forgiveness*, *abuse*, *pride*, etc.). With the help of a pre-compiled emotion lexicon, a keyword spotting approach with a weighted score function may provide an alternative solution to the problem of scarce training samples in emotion classification.

Overall System Architecture

We developed an automated system to detect, at the sentence level, emotion instances from full-text suicide notes. The system architecture is shown in Figure 1. The initial input is a set of full-text suicide

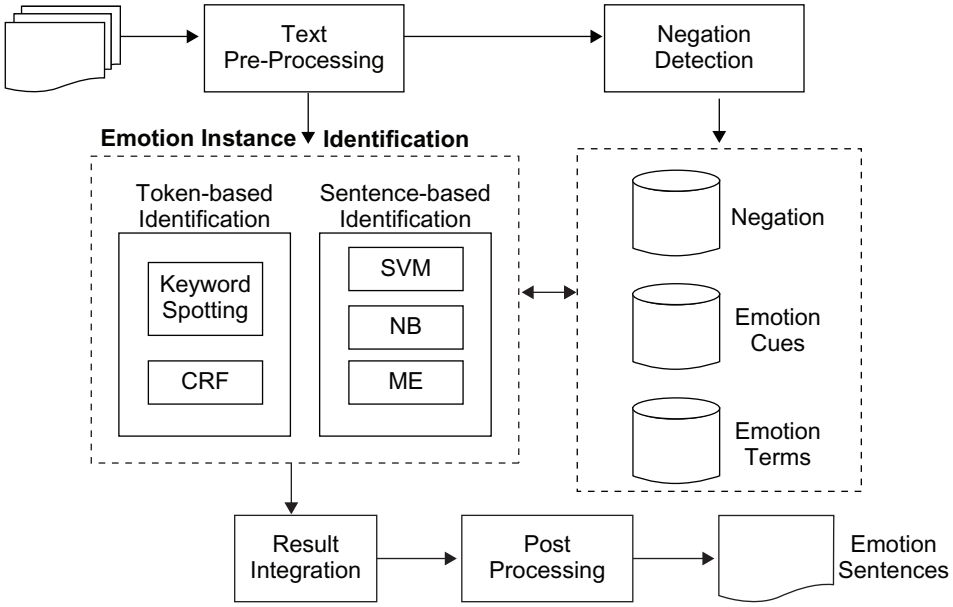

Figure 1. The system architecture diagram.

notes, and the output is the set of selected sentences, each of which contains at least one potential emotion expression and is marked with the corresponding emotion label.

The system consists of five major functional process modules, which are described briefly below:

A. Text pre-processing module

Given a full-text suicide note, it is firstly decomposed into text lines, and each line is treated as a sentence. The individual sentences are processed with the Genia Tagger[9] in order to obtain word lemmas, part-of-speech (POS) tags, and syntactic chunks, and with the Stanford Parser[10] to obtain dependency syntax information used for the machine learner.

B. Negation detection module

This module is used to determine whether the words in a sentence are negation signals (ie, words indicating negation, such as *no*, *not*, *never*). In this system, negation is handled as a modifier to a subject, object, or verb. A hand-crafted lexicon of negation terms are collected from the training data. The lexicon includes 25 words (eg, *no*, *not*, *never*, *unable*) and 38 phrases (eg, *no longer*, *do not*, *have no more*). Negation signals are automatically identified using a simple dictionary look-up method.

C. Emotion instance identification module

This module is designed to identify sentences that explicitly or implicitly refer to a specific emotion state. The module in fact consists of two main components. One is sentence-based identification built on three different machine learning-based models, Support Vector Machine (SVM), Naive Bayes (NB), and Maximum Entropy (ME). The other is token-based identification by a keyword spotting model with a pre-compiled emotion term lexicon and a Conditional Random Field (CRF)-based emotion cue recognition model. For each emotion, each language model first processes the input data and then outputs separate results. These results are combined at the next stage.

D. Result Integration Module

The output results obtained from the different language models are merged together to form an integrated classification result list. Different vote-based merging strategies are used, the details of which are given in the later section.

E. Post-processing Module

This step identifies further instances of the neutral emotions, *information* and *instructions*, which may have been missed by the previous ML models. It applies a number of smoothing rules which recognizes ongoing affective contexts across a number of sentences. The specific rules are given in the later section.

In the following sections, we discuss in detail the behavior of the three important modules, Emotion Instance Identification, Result Integration, and Post-processing.

Emotion Instance Identification

CRF-based emotion cue identification

In this step, we investigate a Conditional Random Field (CRF) model[11] for emotion detection. Given an emotion class to be identified, a sentence is labeled as emotional when it contains some form of emotion cues (ie, the keywords that potentially carry emotion meaning in the sentence).

1. Manual annotation of emotion cues

To construct CRF-based emotion classifiers, we further manually annotated the gold standard of the training data in order to obtain a set of emotion cues for learning. For each sentence marked with one or more emotion class label, we selected emotion fragments from the sentence, ie, the words in the sentence that are impacted by the affect terms. For example, in the example (E5) below, four emotion cues associated with different emotions are selected from the text, "*I love you all*" [Love], "*go to my Mark's Wedding and make him happy*" [Instructions], "*please take care of my darling Bill*" [Instructions], and "*I ca n't go on any more*" [Hopelessness], respectively.

E5: *I love you all! Love, Mary Please, go to my Mark's Wedding and make him happy! Please go! And please take care of my darling Bill, he needs your help now! I hate to do this, but I ca n't go on any more.* [Emotions: *hopelessness, love, and instructions*]

It is noted that each emotion cue is usually made up of at least an affect term (eg, *love*, *go to*, *take care of*, and *ca n't go on*) together with its possible surrounding context words. The annotated emotion instances without any obvious affect term (like the example (E3)) are ignored in the annotation. As a result, we collected a set of 2655 emotion cues from 2173 annotated emotion instances in the training data. This gives a very high coverage of 92.8% against the whole gold standard, where coverage is the proportion of emotion instances which are indicated by one or more emotion cues. The high coverage ratio in cue annotation suggests that most of the emotion events are provoked by some direct or indirect affect terms. This provides strong support for the token-based emotion identification approaches. The cue annotation percentages for different emotion classes are shown in Figure 2. It is interesting that some emotions, such as *hopefulness*, *fear*, *sorrow*, and *anger*, have relative low annotation rates, which implies that underlying semantic emotion expressions frequently appear in the sentences associated with these emotions.

2. Construction of an emotion term lexicon

In constructing our language models, we used a handcrafted lexicon which contains the most salient emotion terms extracted from the training data. Emotion terms are unigrams (eg, *love*), bigrams (eg, *I love*), or trigrams (eg, *I love you*) that convey a particular emotion state. To compile this lexicon, we started with a list of emotion terms which were extracted from the manually-annotated emotion cue set. Then this term list was supplemented by a list of terms that were

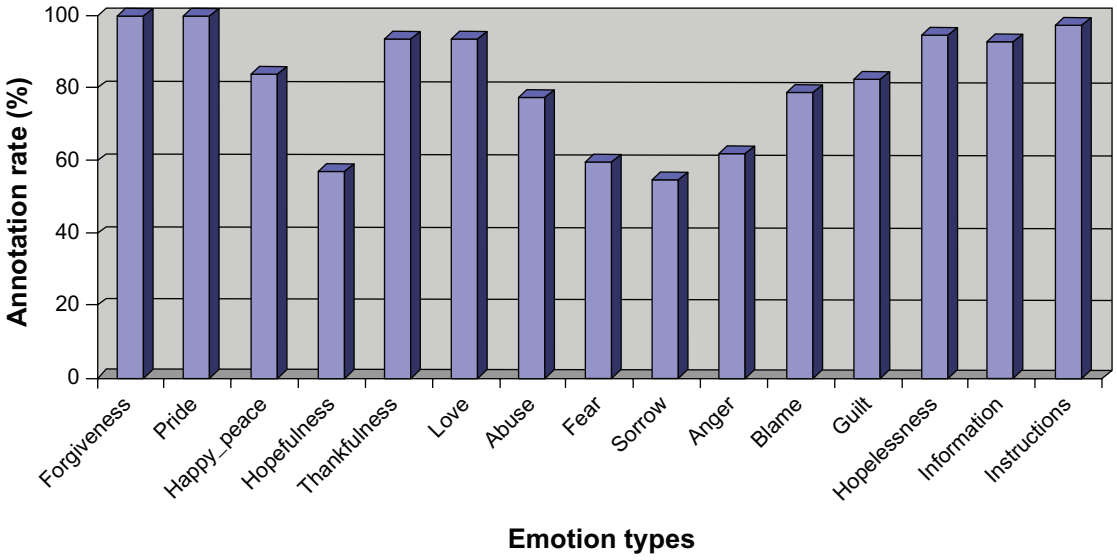

Figure 2. The cue annotation rates for individual emotions.

selected from the annotated emotion instances and were identified as significant by Pearson's chi-square (χ^2) test.[12] We manually checked this complete list and removed those less important terms and finalized a list of 984 emotion terms. Each term is labeled with the referred emotion class and is assigned a weight score that is calculated by the ratio between the number of occurrences in the emotion instances with respect to the specific emotion, and the total frequency in the training data.

3. Features for the CRF-based classification

Each emotion has one CRF-based classifier used for the recognition of emotion cues. We use a wide variety of features to train these CRF-based emotion classifiers. For each of description, we group the features into four sets: word features, context features, syntactic features, and semantic features:

- Word Features: word, word lemma, part-of-speech (POS) tag, phrase chunk tag
- Context Features: 2 previous words and 2 following words with their word lemma, POS tags, and chunk tags
- Syntactic Features: dependency relation label and the governor lemma associated with the word token in focus, which are extracted from the typed dependency information provided by the Stanford Parser.
- Semantic Features: a negation marker that indicates whether this token is a negation signal identified by the negation detection module, and a cue keyword marker that denotes whether this token is a cue keyword in the emotion term lexicon.

We frame the emotion classification task as one of the token-level sequential tagging tasks. Given a sentence, each word token is assigned one of the following tags: B (the beginning of a cue), I (inside a cue), and O (outside of a cue), hereafter referred to as the BIO schema.

4. Sentence labeling

To label the instances of the unseen data we use CRF++[13] to implement our CRF-based language models. Given a sentence, the CRF classifier predicts the presence of emotion cues in the text. If the sentence contains one or more cues with respect to a specific emotion, it will be marked with the corresponding emotion class label.

Lexicon-based keyword spotting

This is the most naive approach, which is to search for the occurrence of particular types of emotion terms in the sentences with the help of the emotion term lexicon discussed earlier. When an emotion term is found in the sentence, the system checks if it is negated by a negation signal. If it is not, add it to a term list associated with the targeted emotion. If one or more emotion terms in terms of a particular emotion are recognized from the sentence, the overall score of the sentence to the emotion is calculated by using a weight score function, ie, the linear combination of all the weights associated with the emotion terms. The sentence is labeled as emotional when the overall score is greater than a weight score threshold τ. Note that the threshold τ for each emotion class is separately set based on the experiments on the training data.

Machine learning-based emotion classification at the sentence level

At the stage of the sentence-level emotion classification, we investigated three different machine learning (ML) algorithms, ie, Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM). The NB and ME language models were implemented by the MALLET toolkit,[14] and the SVM model is trained by the SVM *light*[15] with the linear kernel. We chose these three ML algorithms because they have been proven successful in a number of natural language processing (NLP) tasks such as text classification and Named Entity Recognition (NER), and they represent several different types of learning. We believe that varying the learning algorithms can allow us to obtain more robust and unbiased classification performance by combining the results from different learning algorithms.

A feature vector for a given sentence in the ML-based language models contains the following two sets of features:

- Word Features: word lemma
- Semantic Features: negation terms identified by the negation detection module, and cue terms in the emotion term lexicon

Given an emotion class to be recognized, the feature vector for a sentence is fed into the three different binary classifiers, NB, ME, and SVM, to be distinguished as emotional or non-emotional.

Result Integration

The individual results returned by the different language models are combined using vote-based merging in two stages:

Stage I: the outputs from four different statistical machine learning modules, ie, CRF, NB, ME, and SVM, are combined together according to different voting strategies. The reason that we firstly integrate the results from these four language models is because, unlike the keyword spotting approach that can be applied to all of the 15 emotion classes, the ML-based models merely perform well on six specific emotion classes, *thankfulness*, *love*, *guilt*, *hopelessness*, *information*, and *instructions*, which are provided with enough emotion instances for learning in the training data.

Three different voting strategies have been employed in the result integration of the ML-based models:

- *Any*: if a sentence is identified as an emotion instance by any one of the ML-based models, we consider that it is a true instance of that emotion.
- *Majority*: if a sentence is identified as an emotion instance by two or more of the ML-based models, we consider that it is a true instance of that emotion.
- *Combined*: if a sentence is identified as an emotion instance by two or more of the ML-based models *or* it is identified as an emotion instance by the ML-based model with the best precision for that emotion, we consider that it is a true instance of that emotion.

The *Combined* strategy includes all of the *Majority* integration results plus some of the *Any* integration. The results returned by the specific classifier which has a relative good performance. For example, for the emotion *love*, the CRF model returns 10 annotated emotion sentences that are *NOT* found by other three ML models. Among these annotated sentences, if only 4 are judged as correct by the gold standard, then the accuracy of the CRF model in the *Any* results will be 0.4. However, the NB model provides 8 similar *Any* instances ignored by the other models, but only 1 case is right. Hence, due to the poor performance of the NB model, we only merge the *Any* results by the CRF model into the integrated list.

Stage II: the results from the lexicon-based keyword spotting approach are merged into the above integrated ML-based results in order to form the final classification results.

Post-processing

The post-processing step aims to find more neutral context sentences following the learning stage. This step operates on the observation that when the patients give instructions or other relevant information, they often list a number of items that need to be addressed. The sentences that describe such neutral information are coherent, and thus have some affective continuity. An example of affective continuity is given here.

E5: The neutral-emotion context sentences in the document: 200908031418_1452*ver*2
Line 22: *Mom, all my blankets 8.* [Emotion: *instructions*]
Line 23: *Mary, all dish towels, bath towels etc. 9.* [Emotion: *instructions*]
Line 24: *Jane, all clothes, purses, etc. Sorry, no so hot.* [Emotion: *instructions*]

Two basic smoothing rules are employed in order to find more potential neutral emotion sentences that are missed by the ML-based models during the post-processing:

- Rule 1: If a sentence is not identified as an emotion instance by any ML-based model, but it is between two sentences labeled with the same neutral emotion class, it will be modified as an emotion instance with the same emotion label.
- Rule 2: If a sentence annotated with one of the neutral emotions is followed by an unlabeled sentence, but the unlabeled sentence contains at least one emotion term related to the concerned emotion, the unlabeled sentence will be marked with the same emotion label.

Results

The system was built based on the experiments using 10-fold cross validation over the training set, and system performance reported here was evaluated based on the results of the experiments on the test data. System performance is measured based on recall (R), precision (P), and F-measure (F). Recall is the percentage of the instances correctly against the gold standard. Precision is the percentage of instances classified as affective that are correct in truth. F-measure is the harmonic mean of recall and precision.

Performance of four ML-based language models

To evaluate the performance of four different ML-based language models, CRF, NB, ME, and SVM, we perform a set of experiments on the six major emotions discussed earlier. We compare the classification performance of these four ML-based models. The results of the experiments are given in Table 6. It is noticeable that no one of the learning models stands out as a strong performer. Instead, their performance varies quite widely depending on the different emotion classes. Generally, both CRF model and SVM models perform well in terms of precision, while the NB model excels in recall, and achieves the best micro-average F-measure score with 0.6129. One interesting thing that Table 6 reveals is that the performance of the learning models is not always consistent. A learning model can work well on some specific emotions, but fails on others. For example, compared with the NB and ME model, the SVM model usually achieves good precision but has poor recall. However, for *thankfulness* emotion, it outperforms all of the other three models with a high recall of 0.7067. We observed that compared with other emotions, the emotion keywords frequently occurred in *thankfulness* mostly concentrate on a few specific terms such as *thank*, *thankful*, *appreciate*, *grateful*. One of the possible explanations for the performance of the SVM model is that the SVM model is more sensitive to the frequent context patterns than other three models in emotion identification.

Result integration for four ML-based language models

The inconsistent results obtained by the different ML-based models prompted us to analyze their results, to see whether they compete with or complement one another. Table 7 shows the performance of the integrated results using three different voting strategies: *Any*, *Majority*, and *Combined*. As expected, the *Majority* voting strategy improves precision, while the voting based on positive outcome from the *Any* classifier enhances the overall recall. However, the best merged F-measure is achieved by the *Any* voting method due to the significant improvement in recall with an acceptable precision. The *Combined* voting

Table 6. The performance of the four ML-based models on the test dataset.

	CRF			NB			ME			SVM		
	P	R	F	P	R	F	P	R	F	P	R	F
Thankfulness	0.6056	0.5856	0.5954	0.5881	0.6033	0.5956	0.5778	0.6500	0.6188	0.5263	0.7067	0.6033
Love	0.7620	0.5477	0.6373	0.6850	0.7016	0.6932	0.7762	0.5822	0.6653	0.7687	0.5124	0.6149
Guilt	0.5806	0.2538	0.3532	0.4684	0.3562	0.4046	0.3525	0.2282	0.2770	0.5833	0.2795	0.3779
Hopelessness	0.7353	0.5541	0.6319	0.6326	0.6503	0.6413	0.6867	0.5317	0.5993	0.7451	0.4619	0.5702
Information	0.5314	0.3515	0.4231	0.4096	0.5700	0.4766	0.4811	0.4008	0.4372	0.7619	0.3038	0.4343
Instruction	0.7634	0.5774	0.6574	0.6359	0.6789	0.6567	0.7091	0.5737	0.6342	0.7592	0.4796	0.5878
Micro-average (6 emotions)	0.7150	0.5093	0.5949	0.5998	0.6266	0.6129	0.6528	0.5130	0.5745	0.7228	0.4515	0.5558

Table 7. The merged results after three different integration strategies.

	Any			Majority			Combined		
	P	R	F	P	R	F	P	R	F
Thankfulness	0.5257	0.8156	0.6393	0.6055	0.6567	0.6300	0.5500	0.7711	0.6420
Love	0.6867	0.7563	0.7198	0.7793	0.6399	0.7027	0.7458	0.6767	0.7095
Guilt	0.4655	0.4615	0.4634	0.5000	0.2137	0.2994	0.5052	0.4188	0.4579
Hopelessness	0.6410	0.7081	0.6728	0.8028	0.5522	0.6543	0.6627	0.6838	0.6730
Information	0.4090	0.6485	0.5016	0.4993	0.4181	0.4551	0.4524	0.5073	0.4782
Instruction	0.6849	0.7608	0.7208	0.8079	0.5488	0.6536	0.7032	0.7373	0.7198
Micro-average (6 emotions)	0.6106	0.7067	0.6551	0.7294	0.5195	0.6065	0.6489	0.6597	0.6540

strategy competes with the *Any* classifier in terms of some emotions such as *thankfulness*, *hopelessness*.

Interestingly, in terms of the micro-average performance of the six emotions, the overall F-measure for both *Any* and *Combined* strategies obviously outperforms the best single ML-based model—the NB model. With the combination of the four sets of results, the *Any* F-measure improves by 7.05 points on average across all six emotions compared with the average performance of the individual ML-based models. The substantial improvement in the *Any* classifier suggests that these four ML-based models complement each other very well, and each of them can find some emotion instances that are not predicted by other language models. The integration of the four sets of results allows the system to have a robust and reliable performance.

The overall performance of the system

Our team submitted three runs of results that differ in the choice of result integration strategy and the setting of the weight score thresholds for different emotions in the keyword spotting method. The results of three runs are shown in Table 8. The performance of Run 3 was the best one with a precision of 58.21% and a recall of 64.93%. Nevertheless, the F-measure of Run 3 outperforms the other two runs only by a small margin of less than 1 point. Table 9 reports the detailed evaluation of the performances for the individual emotions. F-measures for the positive emotions range widely from 21.05% (*pride*) to 72.41% (*love*). The negative emotions have a similar wide variety of performances ranging from 20% (*abuse*) to 67.21% (*hopelessness*). The performances for the neutral emotions look better than the negative and positive emotions in which *instructions* emotion has the highest F-measure of 73.3% among all of the emotions.

Interestingly, all of the top performances take place on the six emotions that frequently occur in the dataset, and can be predicted by the four ML-based language models described previously. Compared with the integrated results by the ML-based models introduced in the previous subsection, there is only a slight improvement in F-measure after the results are combined with those from the keyword spotting method and from the post-processing. This suggests that the overall system performance relies heavily on the ML-based language models, with other methods such as the keyword spotting as supplementary.

Figure 3 shows the contribution of different models to the overall system performance. It is obvious that the system performance heavily relies on the effectiveness of the ML-based models. The main reason for this is because the emotion sentences that require to be identified from the main six emotions by the ML-based models account for about 84.7% of all of the emotion instances in the test dataset. It is also observed that the emotions that have infrequent

Table 8. The summary of the evaluation of the three submission runs (Expected—the gold-standard results; Predicted—the results that the system predicted).

	#Expected	#Predicted	#Correct	P	R	F
Run 1	1272	1403	810	0.5780	0.6375	0.6063
Run 2	1272	1560	865	0.5545	0.6803	0.6108
Run 3	1272	1419	826	0.5821	0.6493	0.6139

Table 9. Emotion-based performance of the best submission (Run 3) (Expected—the gold-standard results; Predicted—the results that the system predicted).

	#Expected	#Predicted	#Correct	P	R	F
Positive						
forgiveness	8	4	2	0.5000	0.2500	0.3333
pride	9	10	2	0.2000	0.2222	0.2105
happiness_peacefulness	16	14	8	0.5714	0.5000	0.5333
hopefulness	38	34	10	0.2941	0.2632	0.2778
thankfulness	45	79	39	0.4937	0.8667	0.6290
love	201	205	147	0.7171	0.7313	0.7241
Negative						
abuse	5	5	1	0.2000	0.2000	0.2000
fear	13	16	3	0.1875	0.2308	0.2069
sorrow	34	27	7	0.2593	0.2059	0.2295
anger	26	48	10	0.2083	0.3846	0.2703
blame	45	58	27	0.4655	0.6000	0.5243
guilt	117	123	55	0.4472	0.4701	0.4583
hopelessness	229	259	164	0.6332	0.7162	0.6721
Neutral						
information	104	150	66	0.4400	0.6346	0.5197
instructions	382	390	285	0.7308	0.7461	0.7383

training instances and simply depend on the keyword spotting approach to discover emotion expressions in text have relatively poor performance. This implies that the keyword spotting could not provide the strong discriminative power for emotion identification. This illustrates the limitations of relying on the presence of emotion terms, and the inability of this technique to predict the unseen instances that never appear in the training data.

Discussion

The results reported here demonstrate that an information extraction system can accurately recognize affective text of a variety of emotions involved in suicide notes using natural language processing (NLP) techniques.

In this challenge, statistical machine learning approaches seem to still dominate in emotion identification, and have been proven successful when a large number of manually annotated training instances are available for learning. However, due to the complexity of emotion expressions and ambiguity inherent in natural language, single machine learning algorithm could not provide sustained performance on distinguishing various emotions. As shown in Table 6, the four learning models perform inconsistently over the six emotion classes, which suggests that the characteristics of the six emotions vary widely in emotion expressions, and single ML algorithm has difficulties in dealing with all the differences in emotion expressions. However, when several different learning algorithms work together, the system can perform robustly and provide consistent results.

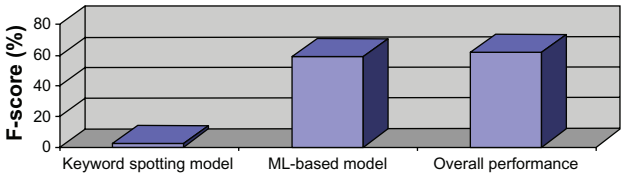

Figure 3. Contribution of different models to the overall performance.

The experimental results in Table 9 show that lexicon-based keyword spotting approach with a weight score function did not perform very well in identifying the emotions with scarce training instances. One of the main causes is that it heavily relies on the occurrence of the emotion terms collected in the emotion term lexicon. The limited coverage of the lexical resource results in the poor recall of the system. Furthermore, token-based keyword spotting might be helpful in sentiment analysis on the basis of local contexts such as words or phrases, but it is not good at handling long-distance emotion expressions. Although we collected a set of bigram and trigram emotion terms that attempt to capture local context information surrounding the affect word, the keyword spotting approach still fails on the detection of emotion expressions, such as the

example (E6), that require an understanding based on the whole clause or sentence.

E6: *I might be able to do something for him.* [Emotion: *hopefulness*]

Sentences like (E6), which contain implicit emotion meaning, account for a large proportion of false negative cases. In our system, both machine learning-based models and the keyword spotting method are incapable of recognizing these sentences that carry affect through underlying meaning, rather than through surface words. Such sentences require a deep semantic analysis of the text. However, a deeper understanding of text is required than what the state-of-art in semantic parsing can provide. How to detect these implicit emotion expressions may be an important avenue for future research on sentiment analysis.

Many false negative and false positive cases are also due to ambiguity in emotion expressions. As emotion is a subjective, a word in similar contexts may provoke different emotions in different people's mind. For example, the sentence (E7) was annotated with the emotion label *pride* because of the affect word "*best*", while the sentence (E8) was recognized as an instance of the emotion *love* evoked by the same affect word. This phenomenon is called *nocuous ambiguity* that occurs when a single linguistic expression is interpreted differently by different people. More discussions about nocuous ambiguity are given in our previous work.[16]

E7: *You are the best wife in the world.* [Emotion: *pride*]

E8: *The best parents anyone ever had.* [Emotion: *love*]

Sometimes, some emotions themselves are ambiguous to each other and hard to distinguish. A typical case is relevant to emotions *blame* and *anger*, which often co-occurred in the text (See Table 5 for frequent co-occurred emotion pairs). Somehow, the ambiguous contexts like (E9) and (E10) lead to inconsistent annotation in the gold standard of both training and test datasets due to different interpretations by the annotators. We consider such ambiguous contexts in fact hint some potential complex interdependencies between different emotions as indicated by the frequent co-occurred emotion pairs. Nevertheless, inconsistent annotation in the gold standard makes our system hard to correctly recognize these ambiguous emotion instances.

E9: *This damn mess my sister has caused has sure and truly been hell.* [Emotion: *blame*]

E10: *My life to you was not worth a damn so maybe by ending it you will be helped.* [Emotion: *anger*]

Related work

Different approaches have already been proposed for textual emotion recognition. Liu et al[17] present a set of commonsense-based linguistic affect models that make use of a knowledge base of commonsense to enable a deep semantic analysis in terms of sentence structure. Mihalcea and Liu employ a corpus-based approach to identify the most salient words for the prediction of the *happy* and *sad* moods in the blogposts. Chaumartin[6] describes a knowledge-based system that investigates a rule-based approach to detect six specific emotions and associated sentiment valence in news headlines with the help of several lexicon resources like WordNet, WordNet-Affect,[18] and SentiWordNet.[19] Masum et al[20] also utilized a rule-based approach to sense emotion from the News by considering cognitive and appraisal structure of emotion and taking into account user preference. Tokuhisa et al[2] propose a two-step approach for the sentence-level emotion classification: first, the sentences are grouped into two categories, *emotion-involved* and *neutral* using a SVM classifier; then, the sentences tagged with *emotion-involved* label are further classified into ten emotion classes by a k-nearest-neighbor (KNN) classifier.

Moreover, a number of researchers work on classifying the contextual polarity of emotion word. Takamura et al[21] use a spin model to extract emotion polarity of words. Quan and Ren[22] explore a variety of features to determine which features are affective for word emotion recognition. Bhowmick and his colleagues[23] propose a transformed network to distinguish emotion words from non-emotion words in WordNet using structural similarity measures.

Our work in textual emotion recognition differs from other research in several ways:

- First, unlike other research that uses publicly available sentiment lexicons such as WordNet, WordNet-Affect, and SentiWordNet, we construct two domain-specific sentiment lexicon resources, emotion cue lexicon and emotion term lexicon, which are directly extracted from the challenge dataset. The domain-specific lexicons provide more reliable emotion clues that appear in the

domain corpus than the public lexicons. These two emotion lexicons have become an important component in our system, and provide strong support for the construction of our five different affect models.

- Second, although the CRF-based approaches have been widely used in the NLP tasks such as token sequential tagging, we believe that this is the first work which introduces CRF-based emotion cue identification for emotion recognition. CRF-based emotion cue identification display evident advantages in sentiment analysis of long, complicated sentences like multi-emotion sentences, because it senses the affect of text on basis of local text fragments other than the whole sentence. The text fragments related to a specific emotion are separately detected, and thus avoid the impact of the noise data concerned with other emotions in a multi-emotion sentence.
- Third, we propose a hybrid model that explores several language models to handle the complicated features inherent in a variety of emotions. Each model has its merit and plays an important role in emotion recognition. These models cooperate and compete to classify the affect of text. The integrated results from different models provide a much more robust and consistent performance.

Conclusion

In this paper we reported on our approach for the 2011 i2b2/VA/Cincinnati Challenge on sentiment analysis in suicide notes. We developed a hybrid model that incorporates several NLP techniques to handle complicated characteristics of affective text related to various emotions involved in suicide notes. Using the domain-specific sentiment lexicons that are constructed directly from the manually-annotated training dataset, the system demonstrates the effectiveness of the proposed hybrid model for automatic emotion recognition with suicide note text. However, the performances in individual emotions suggest that machine learning techniques exhibit a much robust discriminative capability in emotion classification compared with other sentiment techniques such as keyword spotting, especially when a large number of emotions instances are available and when several machine learning algorithms work together and complement to one another. Future work will focus on the detection of the sentences with implicit emotion expressions, and explore methods for effectively identifying the sentences with ambiguous emotions.

Acknowledgements

This paper was awarded as the best research paper in 2011 i2b2/VA/Cincinnati Medical NLP Challenge, Track 2 Shared Task. The work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) as part of the MaTREx project (EP/F068859/1), and by the Science Foundation Ireland (SFI grant 03/CE2/I303_1). The authors wish to acknowledge the anonymous reviewers' useful comments and suggestion, and also would like to thank the challenge organization for organizing this 2011 i2b2/VA/Cincinnati Medical Natural Language Challenge and providing this research opportunity.

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Mihalcea R, Liu H. A corpus based approach to finding happiness. Proc of Computational approaches for analysis of weblogs, AAAI Spring Symposium; 2006.
2. Tokuhisa R, Inui K, Matsumoto Y. Emotion classification using massive examples extracted from the Web. The 22rd International Conference on Computational Linguistics (Coling'08); 2008;881–8.
3. Miller GA. WordNet: A Lexical Database for English. *Communications of the ACM*. 1995;38(11):39–41.
4. Strapparava C, Valitutti A. Wordnet-affect: an affective extension of wordnet. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004); 2004;1083–6.
5. Esuli A, Sebastiani F. Senti-WordNet: A publicly available lexical resource for opinion mining. the Fifth International Conference on Language Resources and Evaluation (LREC 2006); 2006;417–22.
6. Chaumartin FR. A knowledge based system for headline sentiment tagging. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07); 2007;422–5.

7. Li S, Lee SYM, Chen Y, Huang C-R, Zhou G. Sentiment Classification and Polarity Shifting. The 23rd International Conference on Computational Linguistics (COLING'10); 2010;635–43.
8. Ortony A, Clore GL, Collins A. The cognitive structure of emotions. New York: Cambridge University Press; 1988.
9. Genia Tagger: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/.
10. Stanford Parser: http://nlp.stanford.edu/software/lex-parser.shtml.
11. Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the International Conference on Machine Learning (ICML-2001);2001:282–9.
12. Santner J, Duffy DE. The Statistical Analysis of Discrete Data: Springer Verlag; 1989.
13. CRF ++: http://crfpp.sourceforge.net/.
14. The Mallet Toolkit: http://mallet.cs.umass.edu/.
15. SVM *light*: http://svmlight.joachims.org/.
16. Yang H, Roeck Ad, Willis A, Nuseibeh B. A Methodology for Automatic Identification of Nocuous Ambiguity. The 23rd International Conference on Computational Linguistics (Coling'10); 2010:1218–6.
17. Liu H, Lieberman H, Selker T. A model of textual affect sensing using real-world knowledge. Proceedings of the Seventh International Conference on Intelligent User Interfaces (IUI 2003); 2003.
18. WordNet-Affect: http://wndomains.fbk.eu/index.html.
19. SentiWordNet: http://sentiwordnet.isti.cnr.it/.
20. Masum SMA, Prendinger H, Ishizuka M. Emotion Sensitive News Agent: An Approach Towards User Centric Emotion Sensing from the News. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence; 2007;614–20.
21. Takamura H, Inui T, Okumura M. Extracting emotional polarity of words using spin model. the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05); 2005;133–40.
22. Quan C, Ren F. An Exploration of Features for Recognizing Word Emotion. The 23rd International Conference on Computational Linguistics (COLING'10); 2010:922–30.
23. Bhowmick PK, Mukherjee A, Banik A, Mitra P, Basu A. A comparative study of the properties of emotional and non-Emotional words in the Wordnet: A complex network approach. International conference on natural language processing (ICON 2008);2008.